3F3-03

複合現実感アプリケーションのための映像の構造化とその実装

Video Structuring for Mixed Reality Application

堀江 新[*1]　上原 邦昭
Arata Horie　Kuniaki Uehara

[*1]神戸大学大学院自然科学研究科
Kobe University Graduate School of Science and Technology

In this paper, we propose a video structuring method for the mixed reality application. As the video data is serial data, we cannot change the order between two scenes. However, some kind of programs like the news program or the cooking program, we may change order because these video data are sometimes independent from other scenes. Therefore, we use the cooking program and construct the transition model so that we can consider video scenes as parallel steps. Finally we make a mixed reality application that can navigate users with structured video scenes in real time according to requests from users' operations in the kitchen.

1. はじめに

近年，マルチメディア技術の発展により，大量の映像を扱うことが容易になっている．中でも，映像の構造化の研究が盛んに行われている．これは映像を自動的にインデキシングし，ユーザが閲覧しやすくするための研究で，シーンごとに意味を与えて映像全体の構造をとらえるものである．しかしながら，映像の構造化に関する研究は，映像要約や検索に利用されることにとどまっており，他のアプリケーションと統合して活用する事例は少なく，範囲も限定されている．

また，映像の中には料理番組やニュース番組のように，映像の構造が直列的でないものが存在する．たとえば，料理番組は手順に従って映像を流しているが，手順が前後しても問題ないシーンもある．また，ニュース番組では，ニュースを流す順番は制作者側が決定することであり，順序を変更，交換することは可能である．逆に，映画のような映像は，シーンの順番を変えてしまうとストーリーの意味的なつながりがなくなってしまい，映像として見れるものではなくなってしまう．従来の構造化手法では，このようなシーンの交換可能性については考慮されていない．

一方，仮想現実感の研究として，複合現実感という概念がある．複合現実感（Mixed Reality: MR）とは，人間が現実環境からの情報と，PC で処理される下層情報を組み合わせ，相互に利用する技術である．実在するオブジェクトと仮想的なオブジェクトが共に存在する空間を作り出すことができる．本研究では，構造化した映像を用いた複合現実感アプリケーションの構築を検討する．また，シーンの交換可能性，動作特徴に基づいて，並列的な遷移モデルを生成し，調理をするユーザの動作に応じたリアルタイムナビゲーションを可能にする．

2. 複合現実感システムの概要

本研究で提案する複合現実感システムは，料理のリアルタイムナビゲーションシステムである．具体的には，料理番組の映像を利用して，実際に調理をする際に映像で手順をナビゲートできるシステムを構築する．また，複合現実感と料理番組を融合させたシステムの開発を目的としている．

連絡先: 堀江 新 (horie@ai.cs.scitec.kobe-u.ac.jp)
神戸大学　〒657-8501 兵庫県神戸市灘区六甲台町 1-1
Tel: 078-803-6220, Fax: 078-803-6316

本システムでは，机上に MR 環境を実現するために，作業空間となる机を囲むようにしてフレームを組み，机の真上にカメラとプロジェクタを固定している．上部に設置したカメラよりユーザの作業状態を取得し，画像処理によって作業内容の認識を行い，必要な情報をプロジェクタより作業空間へ出力する．作業空間内の実オブジェクトやマーカの認識にはカラー CCD カメラを用いる．また，作業者の手の認識には赤外線カメラを用いる．本システムの構成を図 1 に示す．

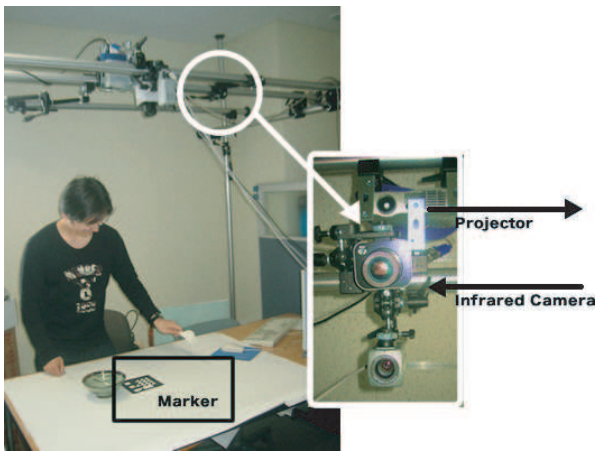

図 1: システムの概要

システムの PC には，構造化された料理映像が蓄積されている．カメラからユーザの動作を解析して，最適な料理番組の映像をナビゲーション映像として表示する．たとえば，チャーハンを作るときを想定する．まず，チャーハンの調理手順を表すモデルをプロジェクタに表示する．ユーザは手順に従って，調理を始めることができる．また，ユーザの「包丁を握る」「火を使う」などの動作を認識し，包丁を使った場合は，利用する具材や切り方などを，火を使う場合は，火の強さや調理時間を作業空間にリアルタイムで表示することが可能である．

システムの入力には，以下のものを利用している．これらの情報を利用して，ユーザが何をしているかを判断し，動作に対応した映像を出力する．

- カラー CCD カメラ

  カメラからの映像はユーザの動作を映す．このカメラか

ら取得される映像を解析すれば，ユーザが，今，何をしているのかを判断することができる．

- 赤外線カメラ

  赤外線カメラによる温度情報を利用すれば，具材の状態の変化やユーザの動作認識ができる．たとえば，作業空間上の温度が高くなっていれば，火を使っていることがわかる．同時に，調理済みの具材も局所的に温度が高くなっていることがわかる．

- マーカトラッキング

  河野らの研究 [1] によるマーカトラッキングは，カメラからマーカを読み取り，そこに蓄えられた情報を獲得できる．本研究では，マーカを机上空間上に置き，情報提供のトリガにすることもできる．利用例として，具材にマーカを配置することを考える．これによって，動作に必要な具材を対応づけることができる．たとえば，包丁を握ったことがわかれば，包丁で切る具材のマーカを点滅させて表示することも可能である．

- ユーザの動作

  寺前らの研究 [2] では，ユーザの手の動きを読み取り，仮想空間上で様々なアクションを起動可能にしている．これを利用して,「レシピを参照する」「もう一度映像を流す」といったアクションを映像配信側にトリガとして送ることも可能である．

以下では，まずプロジェクタに表示する映像を料理番組から構造化するための手法について検討する．料理番組の構造化のステップは,「シーン抽出」,「シーンインデキシング」「遷移モデルの生成」の３ステップに分別される．

3. 映像構造化のためのシーン抽出

映像はカットとショット，シーンを用いて論理的に構造化している．カットとは，カメラが回り始めてから止まるまでの映像の一区切りである．また，ショットとは，カメラワークの開始点や終了点を境界として，カットをさらに細かく分類したものである．ショットは，カメラからの焦点距離の遠近に基づいて，ロングショット，ミドルショット，タイトショットに分類される．複数のカットからなる，映像の意味的な区切りをシーンと呼ぶ．シーンの定義は映像の種類によってあいまいである．そこで，料理番組におけるシーンの区切りと手順の区切りを対応付け，シーンの抽出を自動的に抽出している．

3.1 肌色分布によるショット分類

料理番組において，ショットサイズは，映像の意味を表すうえで非常に有用である．詳細は下記の通りである．

- ロングショット

  料理番組におけるロングショットは，調理人と司会者をとらえるように映される．また，調理中に調理場所の移動が生じるときも必ずロングショットを経由する．

- ミドルショット

  料理番組におけるミドルショットは，調理人にスポットを当てて映されるときに利用される．調理人が料理のポイントを説明するときなどに頻繁に利用される．

- タイトショット

  料理番組におけるタイトショットは，料理中の手元の詳細を映すときに利用される．料理の手順を説明する上で欠かせないショットである．

図 2 に３種類のショットの例を示す．料理番組は，映画のように多様な種類のショットが存在することはなく，これらの定位置から撮影された３種類のショットが切り替わって，シーンとなる．つまり，料理番組中のシーンを区別するためには，まず上の３種類をそれぞれ区別すればよい．

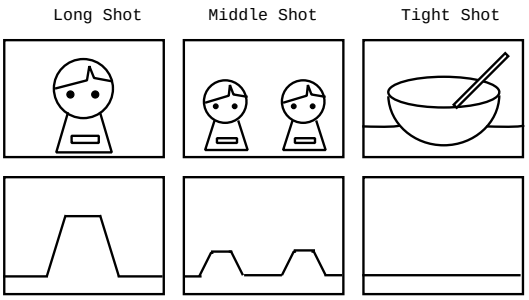

図 2: ３種類のショットと肌色分布

ショットサイズの分類のために，ショット中のキーフレームの画素値を利用している．ショットを区別するうえで有用なパラメータとして，肌色の量を用いている．具体的には，キーフレーム画像における上半分の領域の肌色分布を抽出している．ロングショットの場合，調理人と助手を画面内にとらえるように映像を撮影するために，人物は左右にそれぞれ現れる．そのため，画像中上半分の領域の肌色分布は，左右に小さな山ができる．また，ミドルショットは中央に大きく調理人の顔が現れるため，肌色分布も中央に大きな山ができる．タイトショットは手順によって肌色分布がまったく異なるため，ロングショット，ミドルショットではない分布が現れたときは，タイトショットと判断する．

抽出したタイトショットを前後のタイトショットと比較し，類似度を算出して，シーンの境界を決定している．シーンが変わると，扱う具材や場面も大きく異なる．これは，料理の詳細をもっとも表しているタイトショット間に，大きな変化が生じるためである．そこで，画像的特徴からシーンの境界を発見する．さらに，タイトショット間で大きく類似度が変わった，タイトショットの直前のロングショットを次のシーンの始まりと判断している．

3.2 映像文法を用いたシーン判定

映像の編集を行う際には，編集者の意図を視聴者に伝えようとした場合，ショットのつなぎ合わせ方に規則が存在する．すなわち，映像の概念や事実を伝えるためには，その規則に従う必要がある．この規則の集合を映像文法 [4] と呼ぶ．文献 [3] では，映像文法が映画に利用されていることを前提として，ショットサイズの変化を利用したシーン判定を行っている．ショットサイズの増減を各ショットで抽出し，変曲点をシーン境界として発見している．しかしながら，映像文法に関係する要素はショットサイズのみで，リズムや，カメラワークは境界発見に利用されていない．

本研究では，ショットサイズによる類似度判定に加えて，映像文法の要素の一つでもあるカメラワークをシーン境界抽出に利用している．カメラワークの判定は，投影法 [5] を用いている。投影法は，映像中の輝度変化からカメラワークのパン，

ズームを自動的に取得する手法である．カメラワークを利用すれば，ショット間の類似度に加えて，より正確なシーン判定が可能になる．

- カメラワークがパンの場合

  料理番組において，パンが利用されるときは大きく二種類あり，パンするときのショットサイズによって映像的な意味が変わる．

  - タイトショット時のパン

    タイト時にパンが起きるときは，料理の細部をカメラが追っている状況である．これは番組の冒頭で作る料理を紹介するときや，料理が完成したときに利用されるカメラワークである．

  - ロングショット時のパン

    ロングショット時にパンが起きると，料理の場所が変わるということを表している．たとえば,「包丁で材料を切った後に揚げ物をする」といったことがある．このときは，包丁を切る場所から揚げ物をする場所に，調理人が移動するため，カメラが追うようにパニングが起きるということである．このような区切りはシーンの明確な区切りとなり，シーンの境界として抽出できる．

- カメラワークがズームの場合

  料理番組においてズームが使われるときは，できあがった料理の詳細を見るときが多い．ズームが使用されるタイミングはシーンの途中であることが多く，次のショットでも意味的につながっていると推測される．そのため，ズームが生じているショットでは，シーンが継続していると推測される．

4. 抽出したシーンのインデキシング

抽出したシーンが何の手順を示すシーンか判定するために，文字情報と動作特徴，２つの側面からマッチングを行う．すなわち，料理番組は映像とは別にレシピが記されていることが多い．レシピ中には手順の文字情報が映像と同じ順番で記されている．この文字情報を利用して，シーンの内容を示す手順を発見する．

また，調理の手順には動作の周期性に特徴が存在する．この特徴を利用して手順を推定することが可能である．具体的な手順の特徴を説明する．

- 切る

  「切る」動作には，映像の繰り返しが発生することが多い．これは，手の動きが上下するためであり，映像中のある一カ所に局所的に動きが集中すれば,「切る」と判断することができる．

- 混ぜる

  「混ぜる」動作は,「切る」動作に比べて画面の大部分で動きが発生する．そのため，映像全体で動きが頻発しているのならば,「混ぜる」と判断する．

- 揚げる，焼く

  これらの動作も,「混ぜる」と同様に映像全体で動きが頻発する動作である．しかしながら共にフライパンを利用する動作のため，映像中に黒色が出現する割合が他の動作に比べて高い．そこで，映像中の色分布を調べ，他の動作と区別する．

- 入れる

  ボウルなどに材料を投下する場合，周期的に動きが発生する．また，材料は上から投下することになるため，動きは映像の上半分に集中する．

5. 交換可能性に基づくモデルの自動生成

料理番組のシーン，つまり手順には交換が可能なものがある．同様に，抽出したシーンにも，動作によっては前後関係を交換できるシーンがある．そこで，交換可能性を考慮して，レシピ情報と抽出したシーンの情報を利用した遷移モデルを生成する．図 3 は生成した遷移モデルの具体例である．このモデルは，肉だんごの料理の手順をモデル化したものの一部である．直列的な遷移モデルから，シーンの特徴を利用して並列的な遷移モデルを生成している．

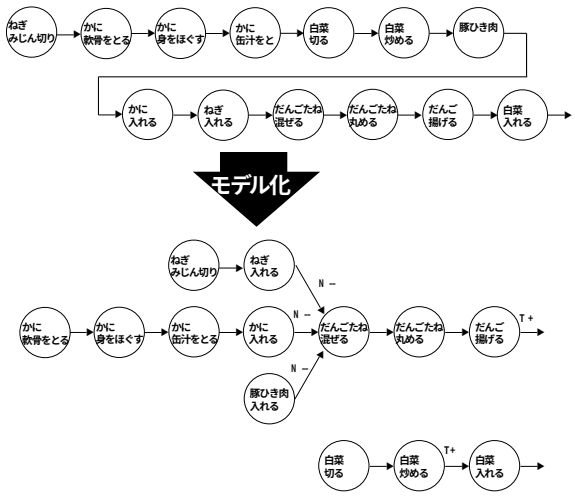

図 3: 遷移モデルの生成の例

5.1 手順の遷移モデル化

手順のモデル化では，まず直列的な遷移モデルを生成するために，レシピをモデル化し，映像情報を付加する．図 3 の上の図はその結果である．レシピに表示された順番で遷移モデルが作成される．しかしながら，実際に料理をするときは，手順はひと通りではなく，複数の料理の手順を並行して行われることも多い．たとえば，野菜類はまとめて切ったり，片方で鍋を使っている間に，もう片方で具材を寝かすなど，実際の料理は並列的に行われる．また，料理番組は，たいてい一番組で2〜3 の料理の手順を紹介する．紹介の方法は直列的で，ひとつの料理を作っている間に別の料理を作るということはない．これも実際に料理をするときとは大きく異なる点で，実際に調理をするときは複数の料理のレシピを考慮に入れた上で最適な手順を作り，調理を行う．

このような映像の直列性に対して，本研究では抽出したシーンの特徴を利用した並列的な遷移モデルを実現し，料理が本来持つ並列性を生み出している．さらに，生成した遷移モデルと，各種カメラを用いたユーザの動作の認識とを比較して，状況に応じたナビゲーションを可能にしている．シーンの特徴を

利用したモデル生成化の例と，ユーザの動作への対応を以下に示す．

- 切る

  図 3 において,「切る」動作は，ねぎ，白菜に出現しているなど,「切る」動作は料理中に複数出現することが多い．同じ「切る」動作の中では，順序に制約は存在しない．ユーザは好きな順番で具材を切ることができるため，この動作は交換可能である．つまり，どちらの手順から始めてもいいように遷移モデルを並列化し，別々の手順とする．並列化することによって，たとえば包丁とねぎを持ったときは，ねぎを切るためのナビゲーションを表示し，包丁と白菜を持ったときは白菜を切るためのナビゲーションを表示するなど，ユーザの動作に対応した映像の表示を可能にする．

- 混ぜる

  「混ぜる」動作は「入れる」動作のあとに続くことが多い．そのため,「入れる」と「混ぜる」は連続して発生すると考えられる．また，混ぜるときには具材がひとつでも欠けてはならない．つまり，図 3 の場合,「だんごたねを混ぜる」動作よりも先に，ねぎ，かに，豚ひき肉に関する動作を終えておかなければならない．そこで，ユーザがこれらの動作を終える以前に混ぜる動作を行おうとしたら，手順を誤っているとしてエラーを返す必要がある．

- 入れる

  「入れる」具材は複数であることが多い．複数の具材を一つにまとめるということは，それ以前の動作を経た具材と，別の動作によってできた具材とを統合するということになる．つまり,「入れる」という動作は，具材をまとめるものである．そのため,「入れる」のあとには具材の数が減少する．さらに，ユーザにはどこに具材を入れるのかをナビゲートする必要がある．

5.2 具材変化の遷移モデル化

調理が進行するにつれ，具材の数や，温度にも変化が生じる．そこで具材に注目して，遷移モデルに情報を加える。まず，具材数は「入れる」手順によって統合されていくため，減少していく．具材の数はマーカを配置すれば認識できる．また，具材の温度は加熱や冷却によって変化する．温度変化は，MR 環境中では赤外線カメラを使用して認識できる．図 3 下の遷移モデルで N-，T+と表記されているものが，それぞれ具材の減少，温度の増加を示している．これらの変化を利用すれば，ユーザの動作をより正確に認識することが可能である．たとえば，具材の数が減少すれば「入れる」の動作が完了したというトリガになり，ユーザの動作が温度変化を伴っていれば,「揚げる」や「焼く」動作であると認識できる．

6. 関連研究

料理番組の構造化にはさまざまなアプローチがある．大きく分けて，画像処理によるアプローチと文字情報によるアプローチの２種類がある．本稿で提案した手法では，映像情報から抽出したシーンにレシピの情報を対応づけ，構造化している．他に進められている研究として，文字情報として番組中のクローズドキャプションを利用し，映像情報を音声解析，画像解析した結果をテキスト教材の文字情報と対応づける研究 [6] や，クローズドキャプションの発話構造を解析して，階層構造を構築する研究 [7] がある．これらの構造化は料理映像の検索システムや，映像要約システムに応用されているが，映像を直列的に解釈しており，交換可能性については述べられていない．また，構文解析をする際には，大量の言語情報を辞書に入力する必要があり，処理に時間がかかるという問題もある．

また，レシピ中の動作をまとめた辞書を構築し，代表的な動作をアニメーションに変換する研究が進められている [8]．この研究では，映像情報を解析には利用していないが，文字情報であるレシピを解析して新たな映像情報を作りだしている．

7. まとめ

本稿では，料理番組を映像を構造化するために，画像類似度とカメラワークからシーン境界を自動で抽出し，テキスト情報とマッチングを行った．また，複合現実感と組み合わせて，料理番組を用いたユーザの動作と対応可能なリアルタイムナビゲーションシステムの実現について検討した．今後は，システムの実装を目指すと共に，映像編集面でも複数の料理を入力とした場合においても，自動で構造化し再編集できるように応用していくつもりである．

参考文献

[1] 河野 武, 伴 好弘, 上原 邦昭, “ウェアラブルコンピュータのためのビデオトラッキング用コード化マーカについての検討”, 日本バーチャルリアリティ学会論文誌, Vol.8, No.3, pp.311-320 (2003).

[2] 寺前 雄亮, 伴 好弘, 上原 邦昭, “複合現実感を利用した机上コラボレーションシステムの開発”, 電子情報通信学会技術研究報告, vol.104, No.572, PRMU2004, pp.1-6 (2005).

[3] Jihua Wang, and Tat-Seng Chua, “A Framework for Video Scene Boundary Detection”, Proc. of ACM Multimedia 2003, pp.243-246 (2003).

[4] ダニエル アリホン, “映画の文法”, 紀伊國屋書店 (1980).

[5] 長坂 晃朗, 宮武 孝文, “輝度投影空間を用いたビデオモザイク”, 電子情報通信学会論文誌, Vol.J-82-II, No.10, pp.1572-1580 (1999).

[6] 三浦 宏一, 高野 求, 浜田 玲子, 井出 一郎, 坂井 修一, 田中 英彦, “料理映像の構造解析による調理手順の対応付け”, 電子情報通信学会論文誌, Vol.J-86-II, No.11, pp.1647-1656 (2003).

[7] 柴田 知秀, 黒橋 禎夫, “料理教示発話の理解と作業構造の自動抽出”, 情報処理学会「自然言語処理」研究報告, Vol.2004, No.93, pp.117-122 (2004).

[8] 白井 清昭, 大川 寛志, “アニメーション生成のための料理動作辞書の構築”, 情報処理学会「自然言語処理」研究報告, Vol.2004, No.93, pp.123-128 (2004).